PACE UNIVERSITY

HACKATHON 2021

THEME : EXPLAINABLE AI

TEAM # : REVENUE REVEALERS

"Your **DATA** (mind) is like this water, my friend. When it is agitated, it becomes difficult to see. But if you allow it to settle, the answer becomes clear."

Presented by Ashika (user_34), Zeeshan(user_9), Ankit (user_8) & Raviraj(user_7)

Executive Summary

- RESEARCH PROBLEM
- STRATEGIC PLANNING FOR PRICE PREDICTION
- EXPLAINING AI / ML STRATEGY TO ACHIEVE PRICE PREDICTION
- PRICE PREDICTION
- INTERPRET TO FIND ANSWERS & CONCLUSION

Research Problem

Data 1

2 Tool

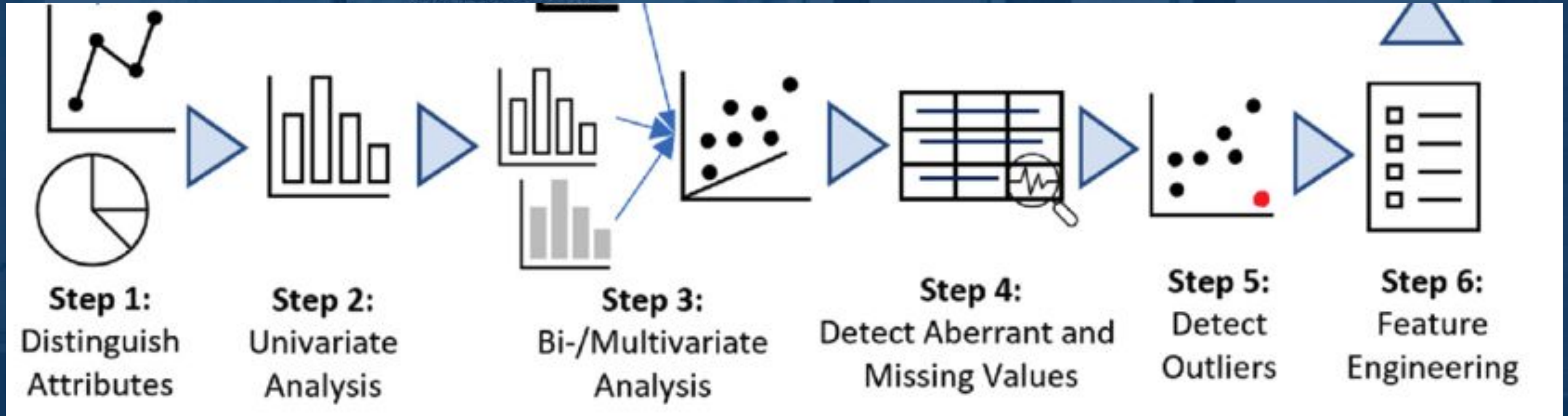
AI / ML TECHNOLOGY

Buyer 3

4 Seller/ Agent

Strategic Planning for Price Prediction

EXPLORATORY DATA ANALYSIS - ZILLOW CLEANED



Step 1: Distinguish Attributes

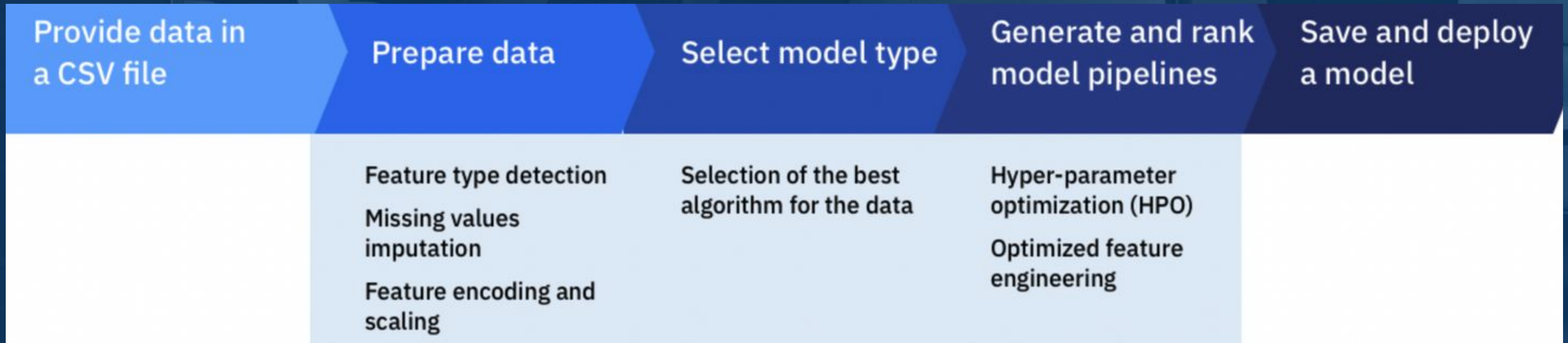
Step 2: Univariate Analysis

Step 3: Bi-/Multivariate Analysis

Step 4: Detect Aberrant and Missing Values

Step 5: Detect Outliers

Step 6: Feature Engineering



Provide data in a CSV file

Prepare data
- Feature type detection
- Missing values imputation
- Feature encoding and scaling

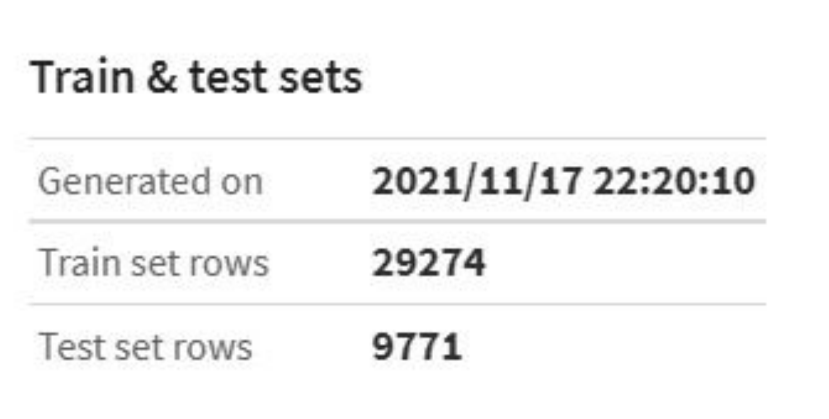
Select model type
- Selection of the best algorithm for the data

Generate and rank model pipelines
- Hyper-parameter optimization (HPO)
- Optimized feature engineering

Save and deploy a model

APPROACH TO PREDICT PRICE

- **Split data into**
  - ○ **Train - 0.75**
  - ○ **Test - 0.25**



Train & test sets

Generated on	2021/11/17 22:20:10
Train set rows	29274
Test set rows	9771

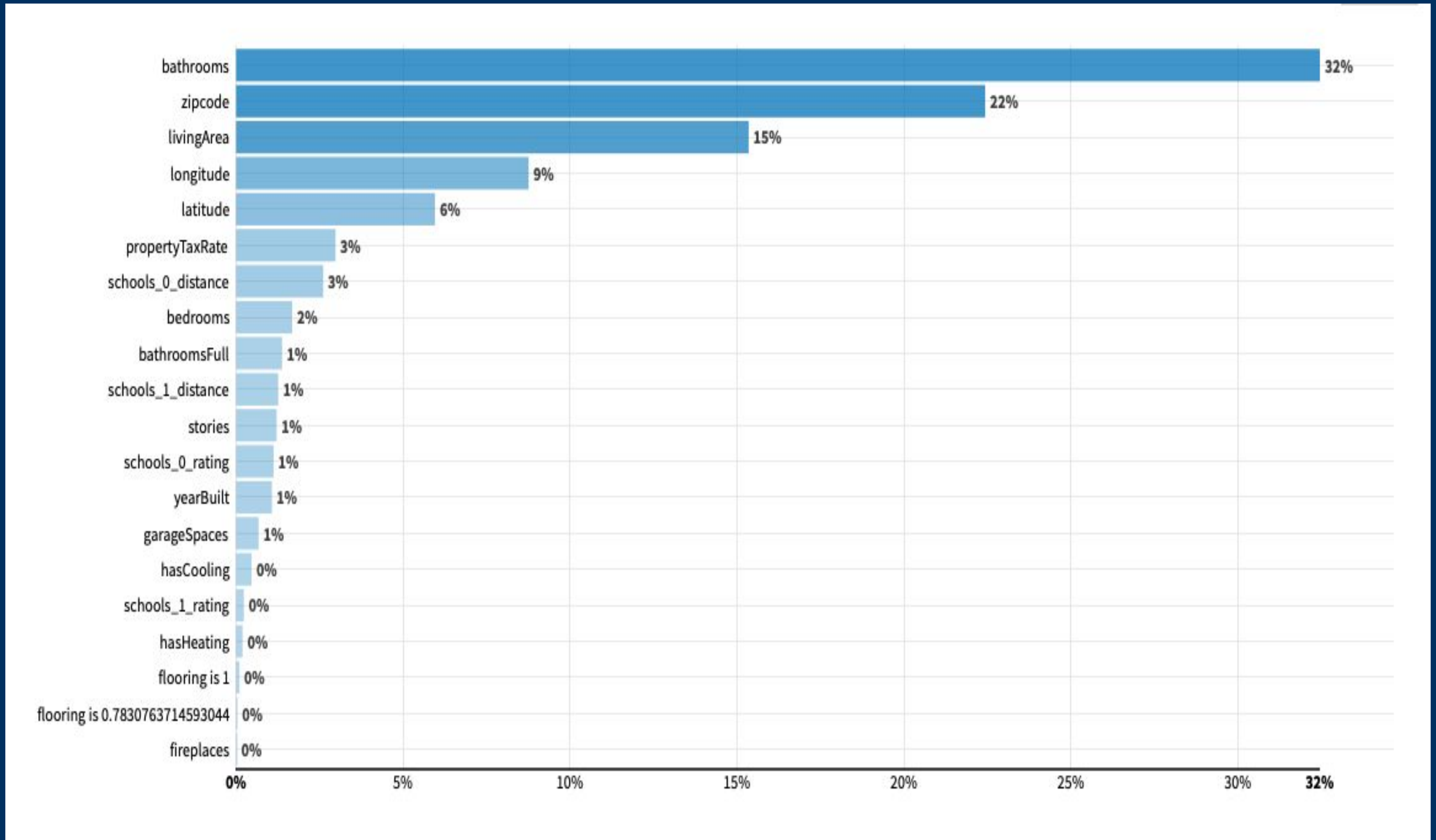
- **Optimal feature selection to obtain the best outcome.**

Explaining AI / ML Strategy for Price Prediction

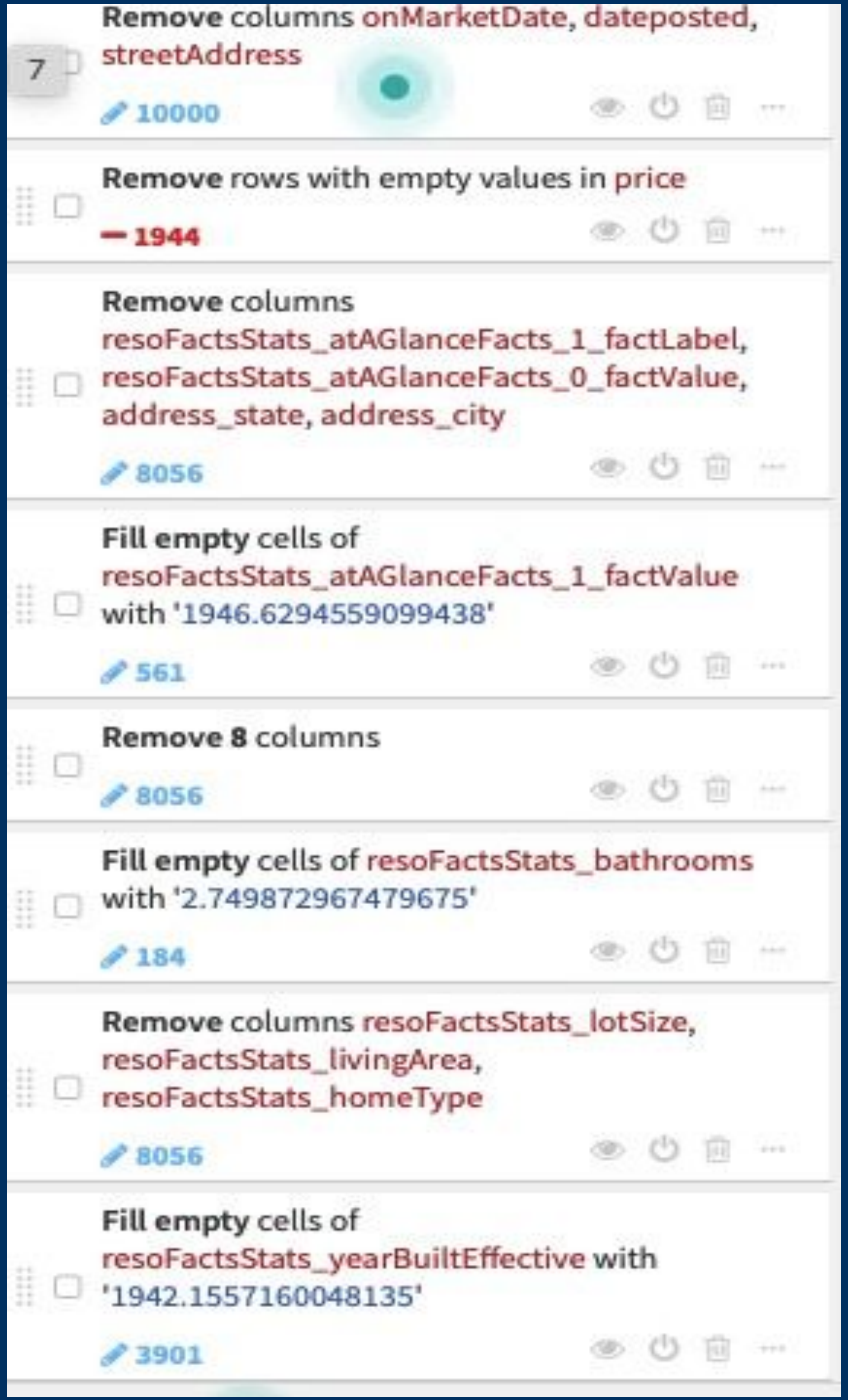
Exploratory Data Analysis is an approach/philosophy of an Data Analysis that employs a variety of techniques of.

- maximize insight into a data set
- uncover underlying structure
- extract important variables*
- detect outliers and anomalies
- test underlying assumptions
- develop parsimonious models

Plotting the raw data (such as data traces, histograms, bihistograms, probability plots, lag plots, block plots, and Youden plots.



Data Preprocessing & Feature Tuning



7 □ **Remove** columns onMarketDate, dateposted, streetAddress
10000

□ **Remove** rows with empty values in price
— 1944

□ **Remove** columns resoFactsStats_atAGlanceFacts_1_factLabel, resoFactsStats_atAGlanceFacts_0_factValue, address_state, address_city
8056

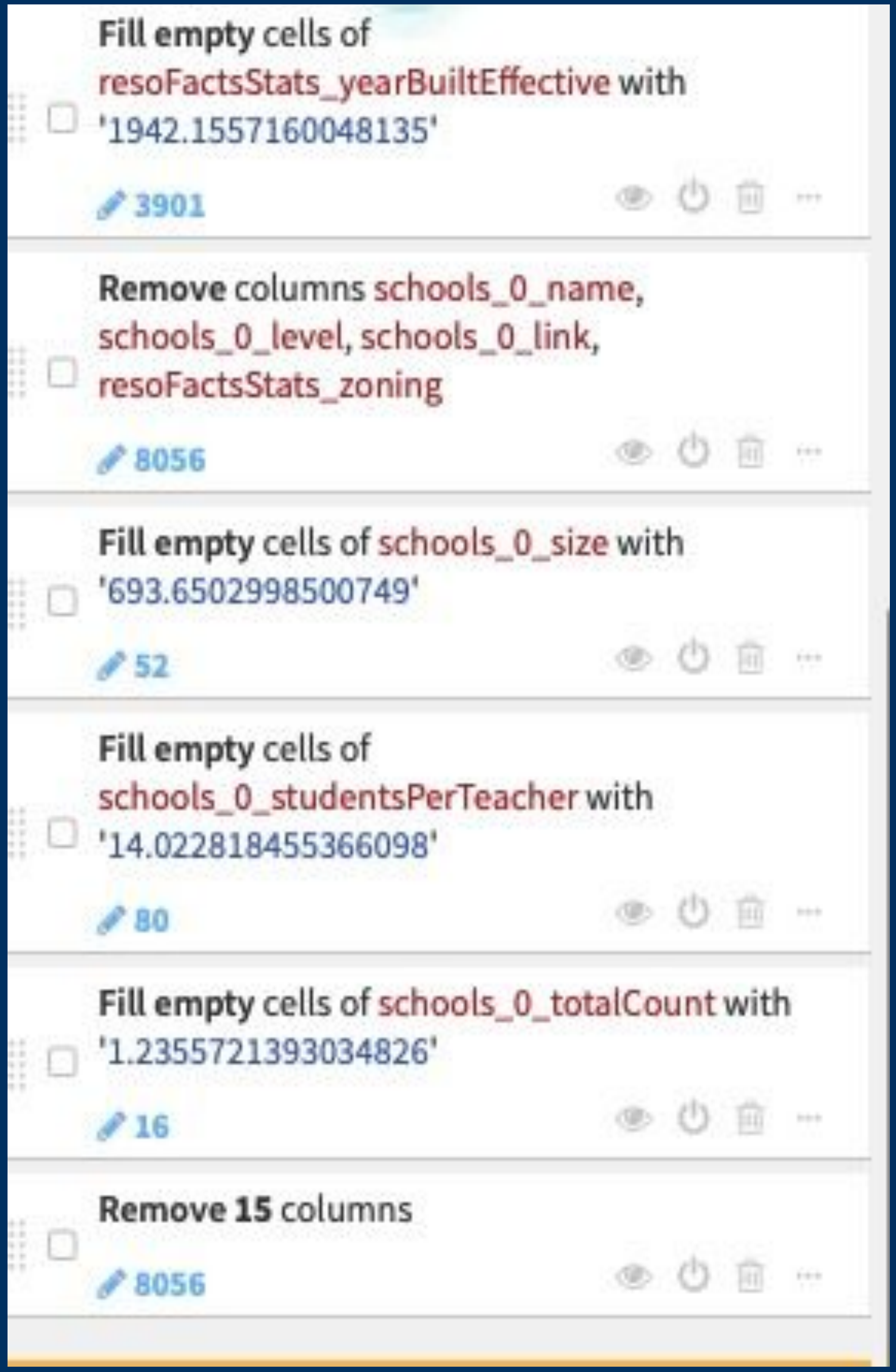
□ **Fill empty** cells of resoFactsStats_atAGlanceFacts_1_factValue with '1946.6294559099438'
561

□ **Remove 8** columns
8056

□ **Fill empty** cells of resoFactsStats_bathrooms with '2.749872967479675'
184

□ **Remove** columns resoFactsStats_lotSize, resoFactsStats_livingArea, resoFactsStats_homeType
8056

□ **Fill empty** cells of resoFactsStats_yearBuiltEffective with '1942.1557160048135'
3901



□ **Fill empty** cells of resoFactsStats_yearBuiltEffective with '1942.1557160048135'
3901

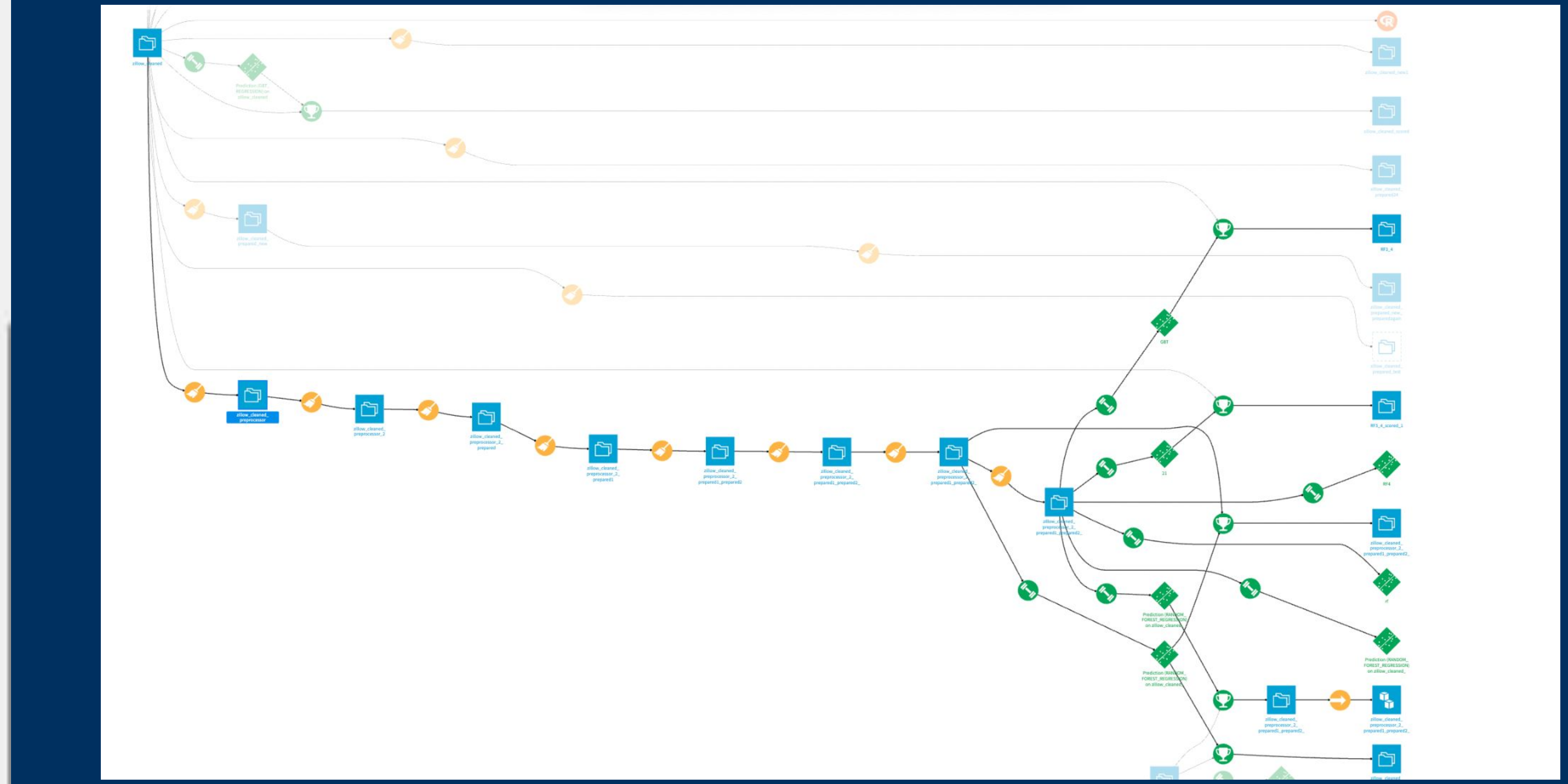
□ **Remove** columns schools_0_name, schools_0_level, schools_0_link, resoFactsStats_zoning
8056

□ **Fill empty** cells of schools_0_size with '693.6502998500749'
52

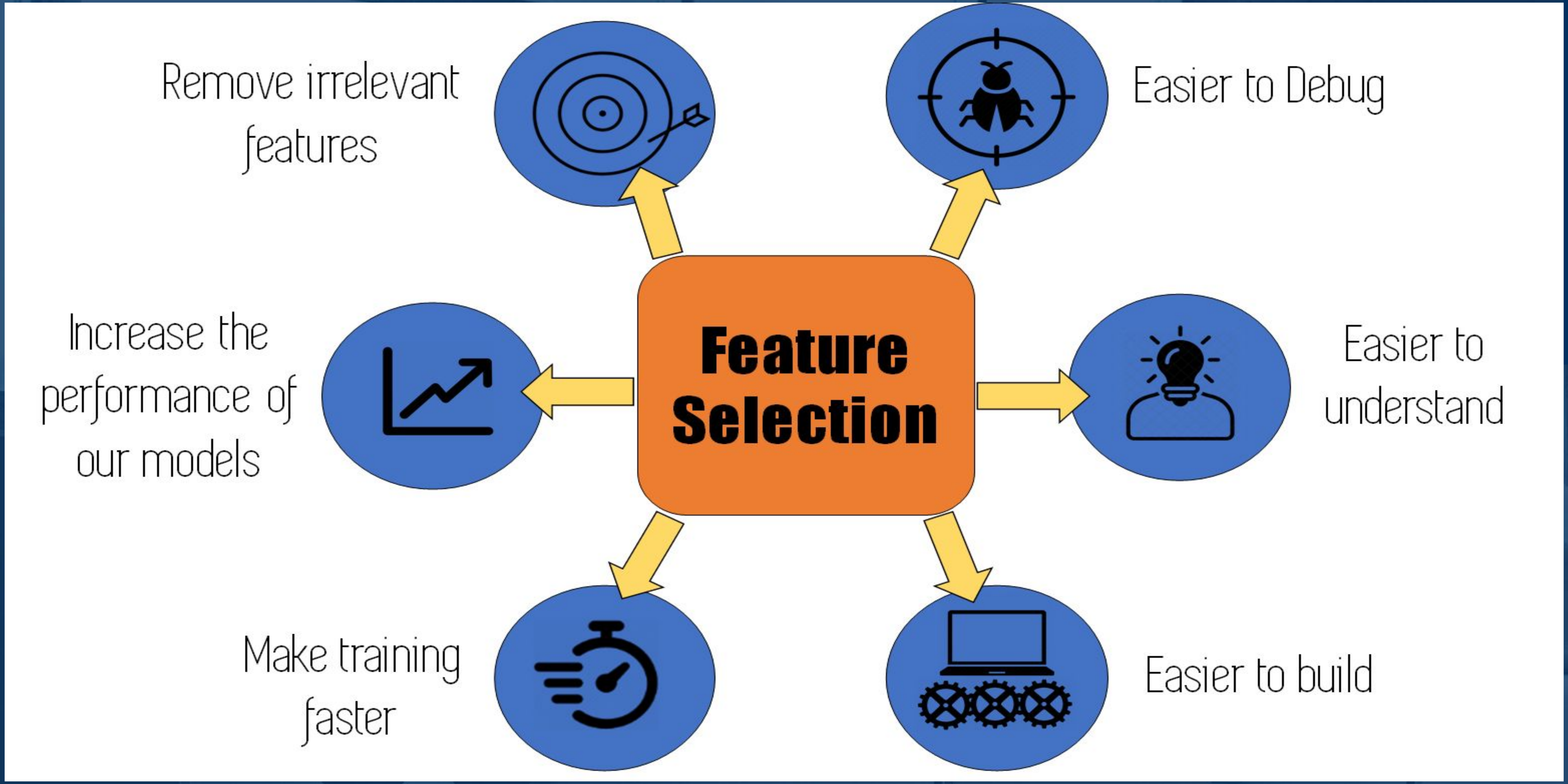
□ **Fill empty** cells of schools_0_studentsPerTeacher with '14.022818455366098'
80

□ **Fill empty** cells of schools_0_totalCount with '1.2355721393034826'
16

□ **Remove 15** columns
8056



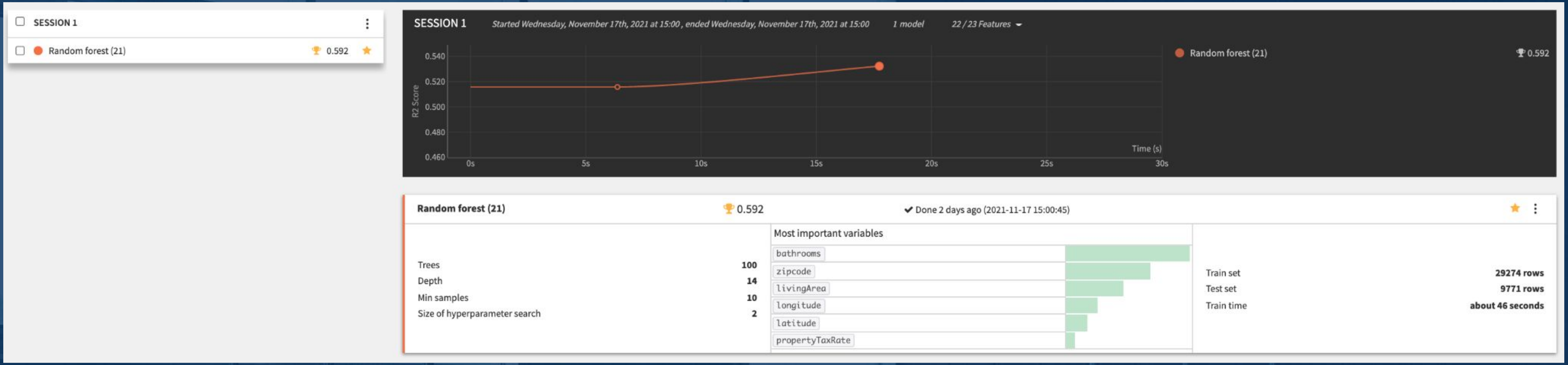
Feature Engineering



BATHROOMS
ZIP CODE
LATITUDE
LONGITUDE
LIVING AREA
PROPERTY TAX RATE

PRICE_HISTORY _01
PRICE_HISTORY _02
.
.
.
.
PRICE_HISTORY _29

Explaining AI / ML Strategy for Price Prediction



□ SESSION 1 ⋮

□ ● Random forest (21) 🏆 0.592 ★

SESSION 1 *Started Wednesday, November 17th, 2021 at 15:00 , ended Wednesday, November 17th, 2021 at 15:00* *1 model* *22 / 23 Features*

Random forest (21) 🏆 0.592 ✓ Done 2 days ago (2021-11-17 15:00:45) ★ ⋮

Trees	100
Depth	14
Min samples	10
Size of hyperparameter search	2

Most important variables

- bathrooms
- zipcode
- livingArea
- longitude
- latitude
- propertyTaxRate

Train set	29274 rows
Test set	9771 rows
Train time	about 46 seconds

KEY FINDINGS

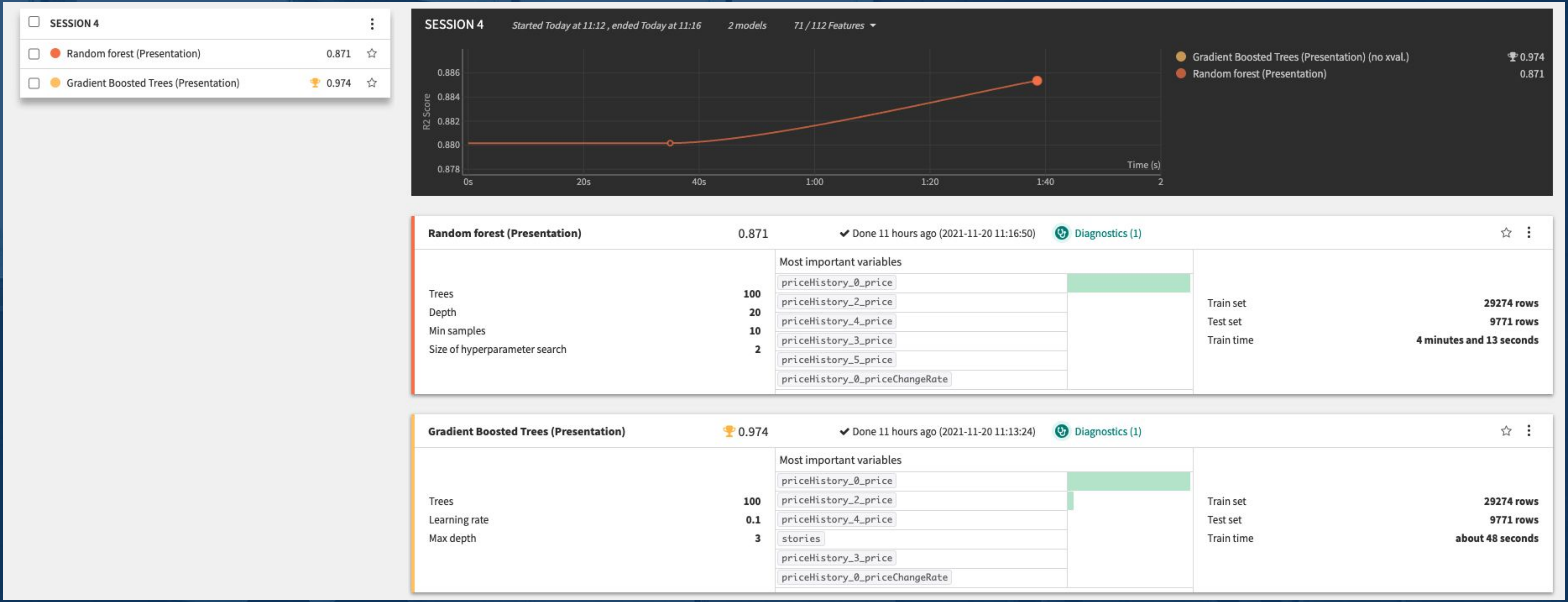
- Initial Slide provides us with an details on importance of certain feature while we try to generate the best model
- Removal text features which played no major role in better model generation
- Price_history columns from raw was an addition to zillow_cleaned data for better results

PRICE_HISTORY _01 ,PRICE_HISTORY _02, ……. PRICE_HISTORY _29

CONCLUSION

- Best Model Generated.
- Better Prediction Results.

Explaining AI / ML Strategy for Price Prediction



SESSION 4

□	● Random forest (Presentation)	0.871	☆
□	● Gradient Boosted Trees (Presentation)	🏆 0.974	☆

SESSION 4 *Started Today at 11:12 , ended Today at 11:16* *2 models* *71 / 112 Features*

● Gradient Boosted Trees (Presentation) (no xval.)	🏆 0.974
● Random forest (Presentation)	0.871

Random forest (Presentation) 0.871 ✔ Done 11 hours ago (2021-11-20 11:16:50) Diagnostics (1)

Trees	100
Depth	20
Min samples	10
Size of hyperparameter search	2

Most important variables

- priceHistory_0_price
- priceHistory_2_price
- priceHistory_4_price
- priceHistory_3_price
- priceHistory_5_price
- priceHistory_0_priceChangeRate

Train set	29274 rows
Test set	9771 rows
Train time	4 minutes and 13 seconds

Gradient Boosted Trees (Presentation) 🏆 0.974 ✔ Done 11 hours ago (2021-11-20 11:13:24) Diagnostics (1)

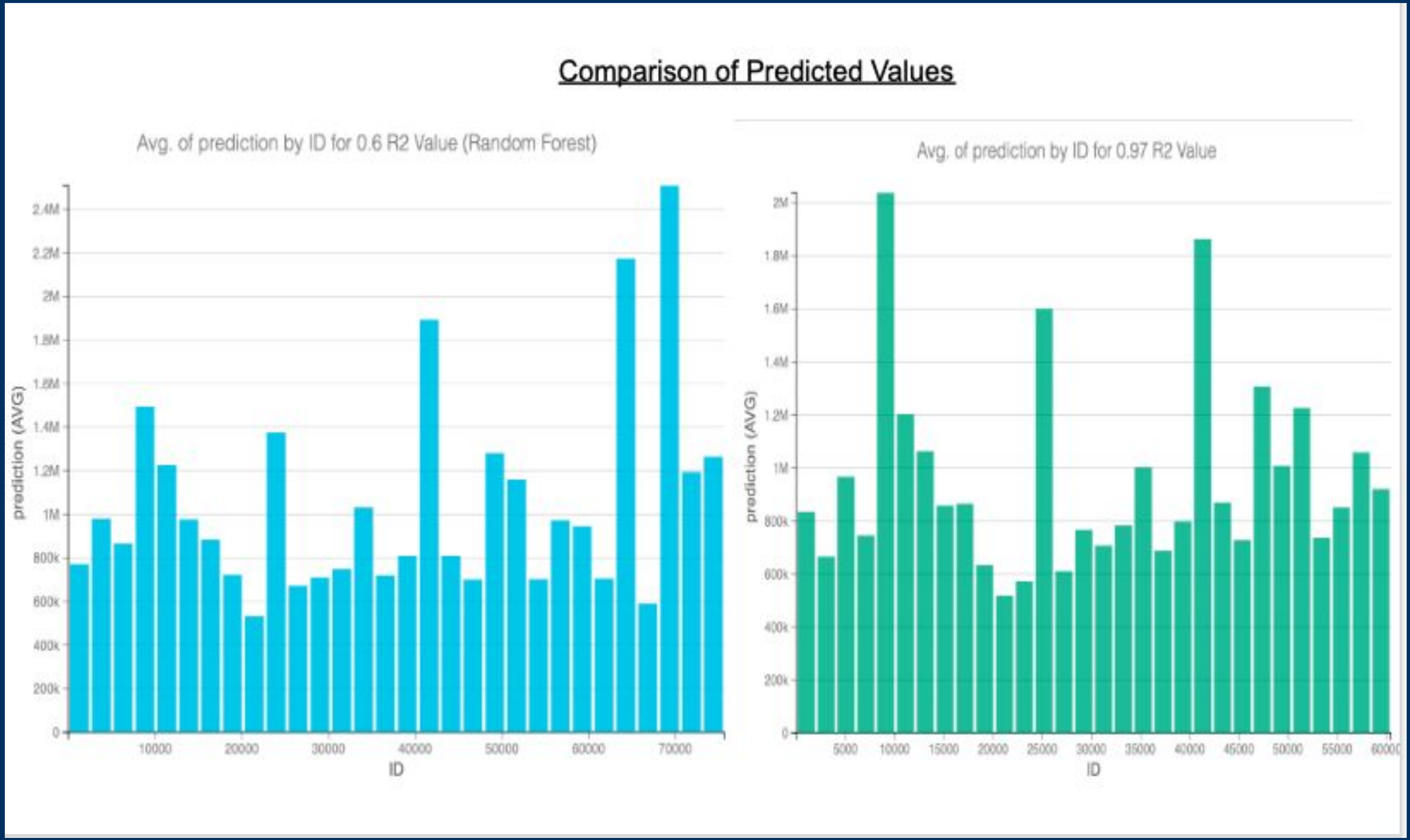
Trees	100
Learning rate	0.1
Max depth	3

Most important variables

- priceHistory_0_price
- priceHistory_2_price
- priceHistory_4_price
- stories
- priceHistory_3_price
- priceHistory_0_priceChangeRate

Train set	29274 rows
Test set	9771 rows
Train time	about 48 seconds

Data Interpretation

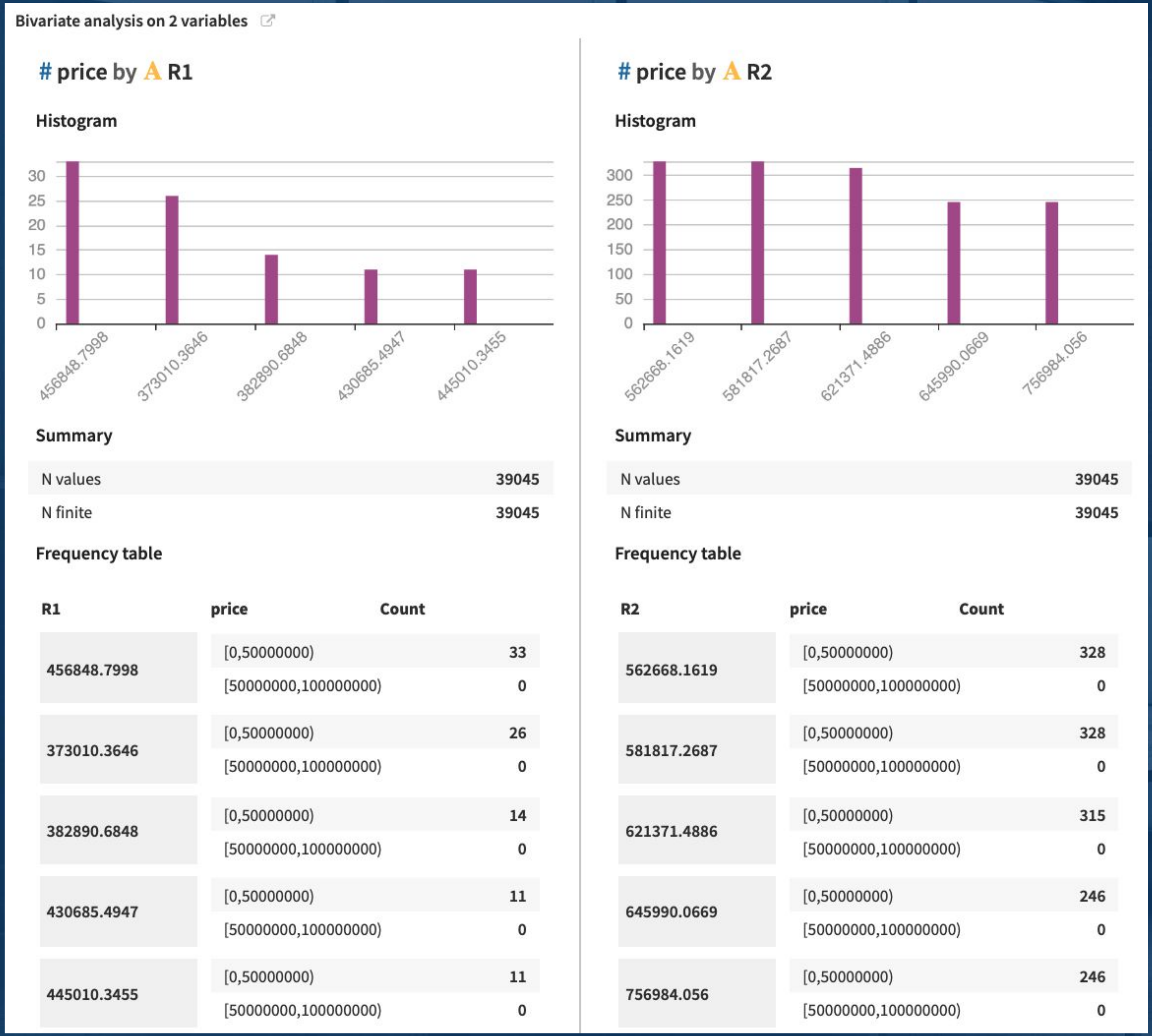


- Comparison of Predicted Values before and after addition of price_History columns in the dataset.

- The zillow raw which is an external source was used in order to obtain extra features for better R2 score and better predictions.

- Even though the predicted R2 model for zillow raw had XGBOOST had highest R2 score but GBT provided us with better prediction results

Comparison of AI / ML Strategy for Price Prediction



Bivariate analysis on 2 variables

price by A R1

Histogram

Summary

N values	39045
N finite	39045

Frequency table

R1	price	Count
456848.7998	[0,50000000)	33
	[50000000,100000000)	0
373010.3646	[0,50000000)	26
	[50000000,100000000)	0
382890.6848	[0,50000000)	14
	[50000000,100000000)	0
430685.4947	[0,50000000)	11
	[50000000,100000000)	0
445010.3455	[0,50000000)	11
	[50000000,100000000)	0

price by A R2

Histogram

Summary

N values	39045
N finite	39045

Frequency table

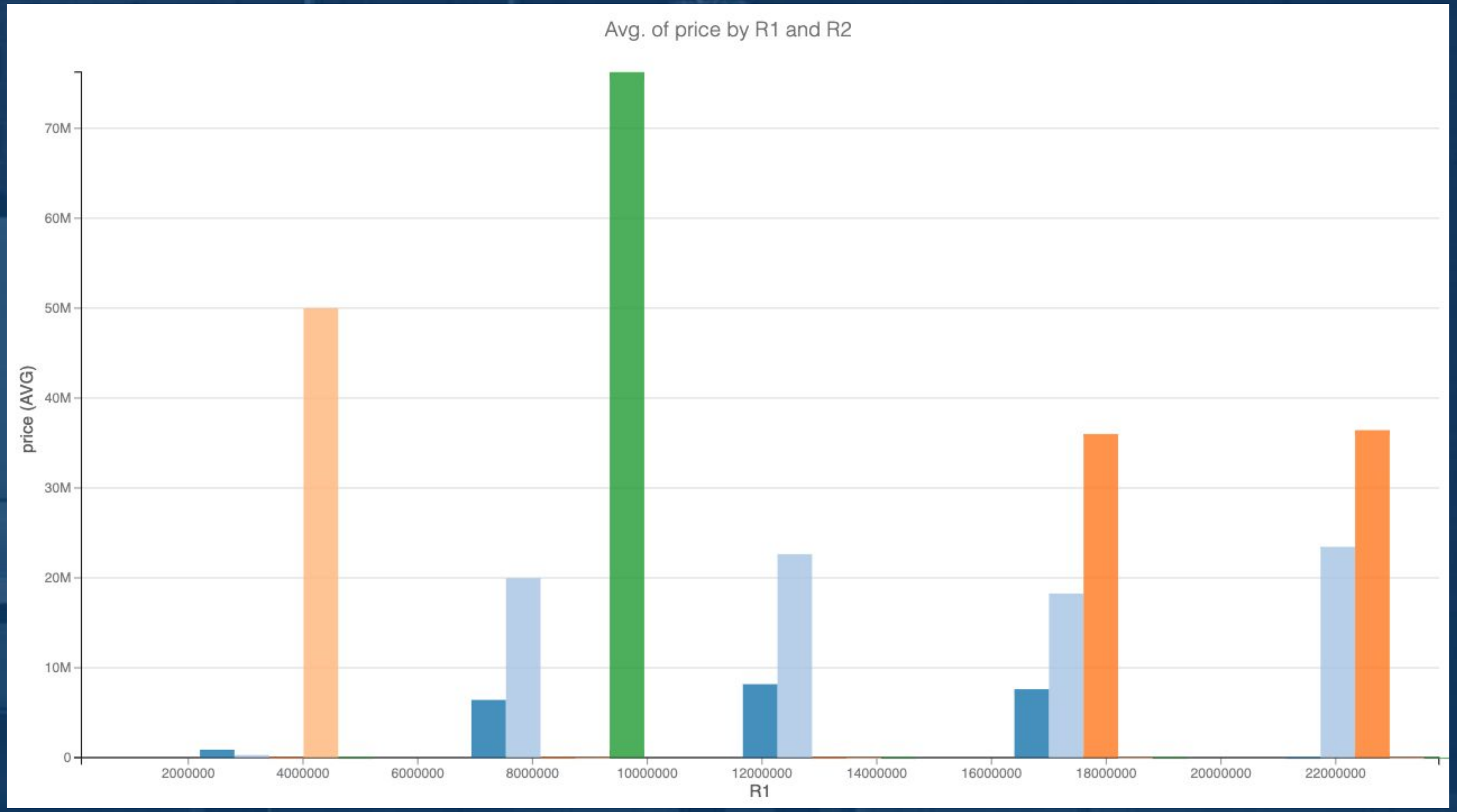
R2	price	Count
562668.1619	[0,50000000)	328
	[50000000,100000000)	0
581817.2687	[0,50000000)	328
	[50000000,100000000)	0
621371.4886	[0,50000000)	315
	[50000000,100000000)	0
645990.0669	[0,50000000)	246
	[50000000,100000000)	0
756984.056	[0,50000000)	246
	[50000000,100000000)	0

ZILLOW CLEANED DATA

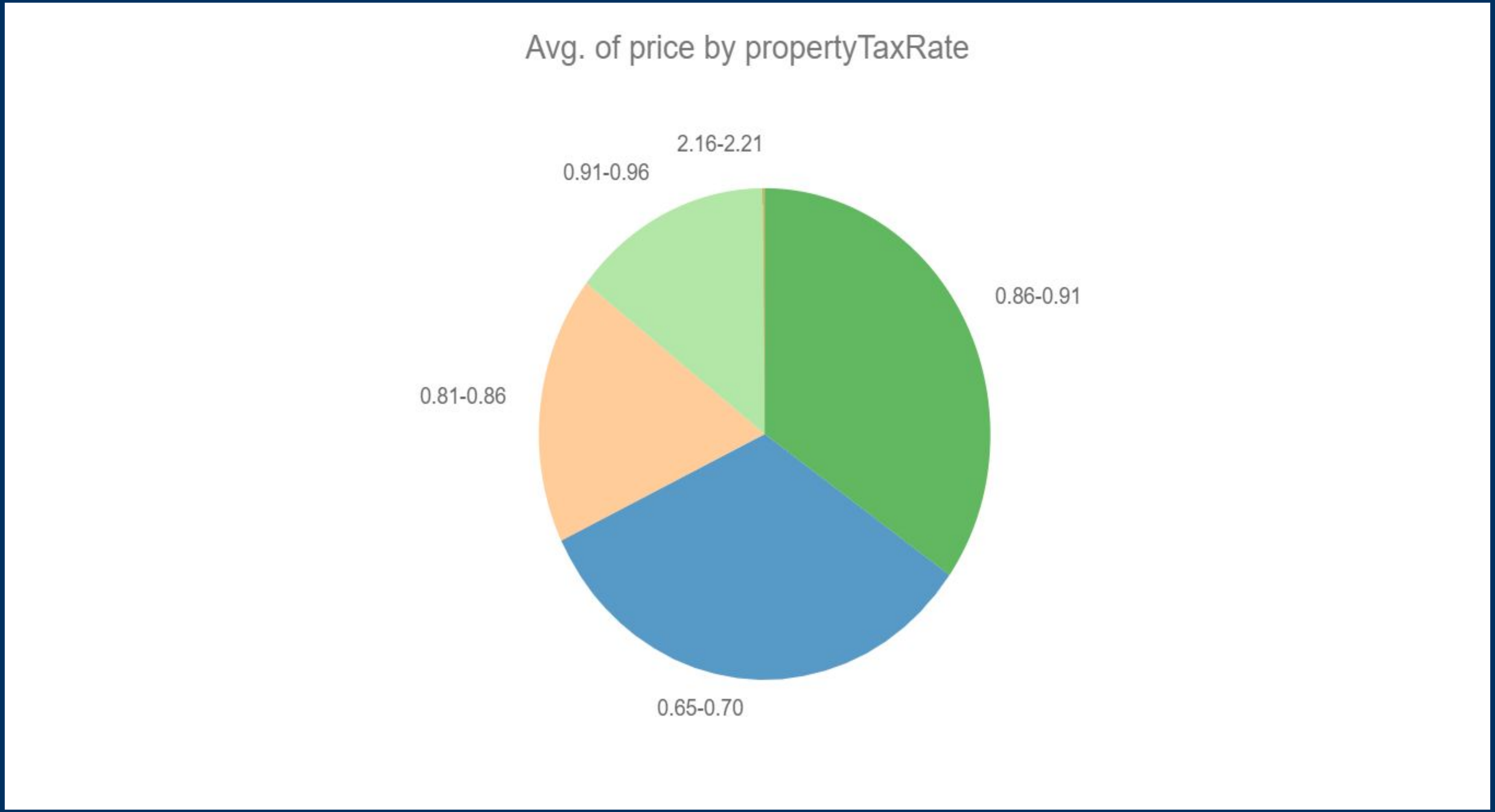
0.59

ZILLOW RAW DATA

0.97



Data Interpretation



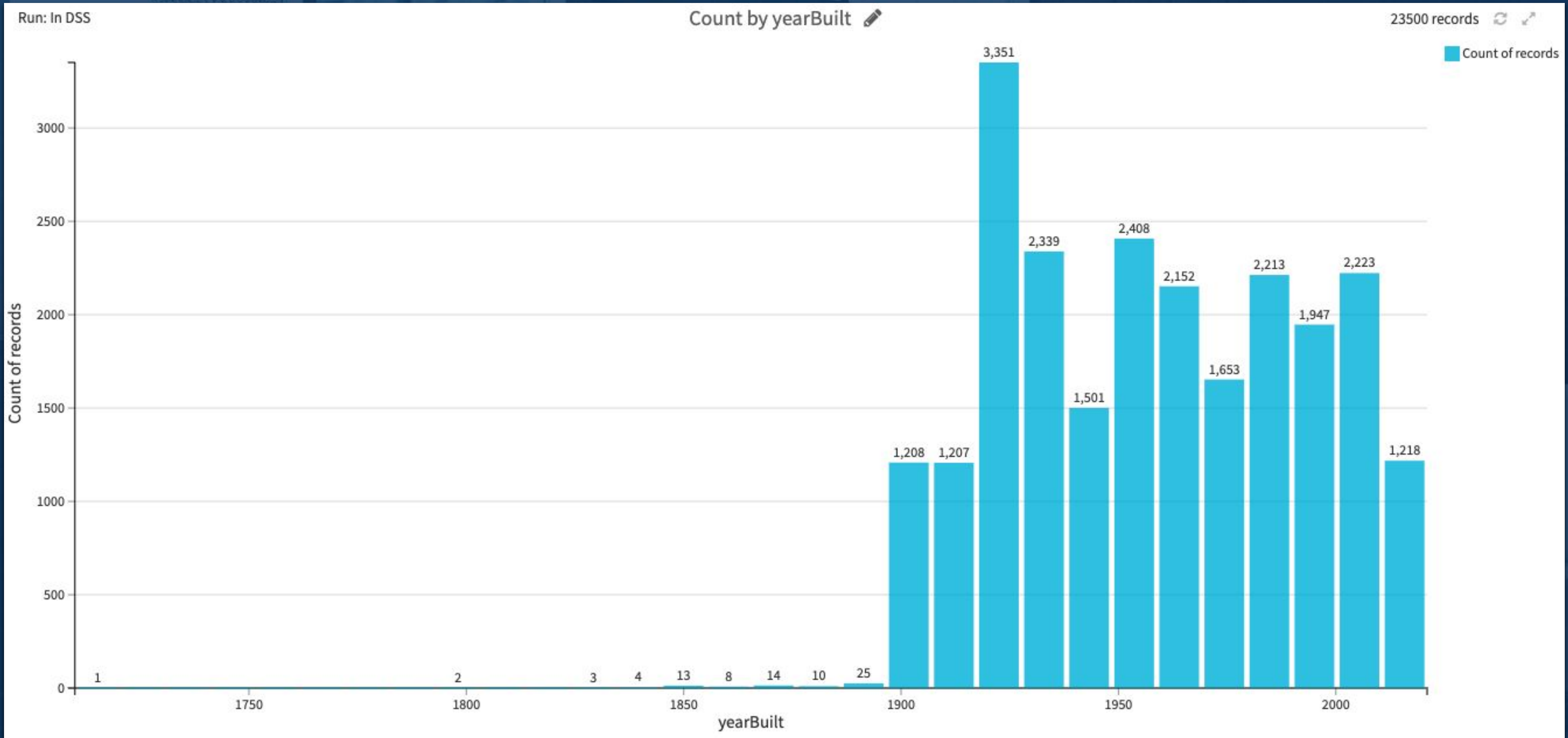
- **The Pie chart represents avg price by property TaxRate.**

- **The highest average of price is in the range of 0.86-0.91**

- **Quantifying the range to be great for a highest pricing of the house**

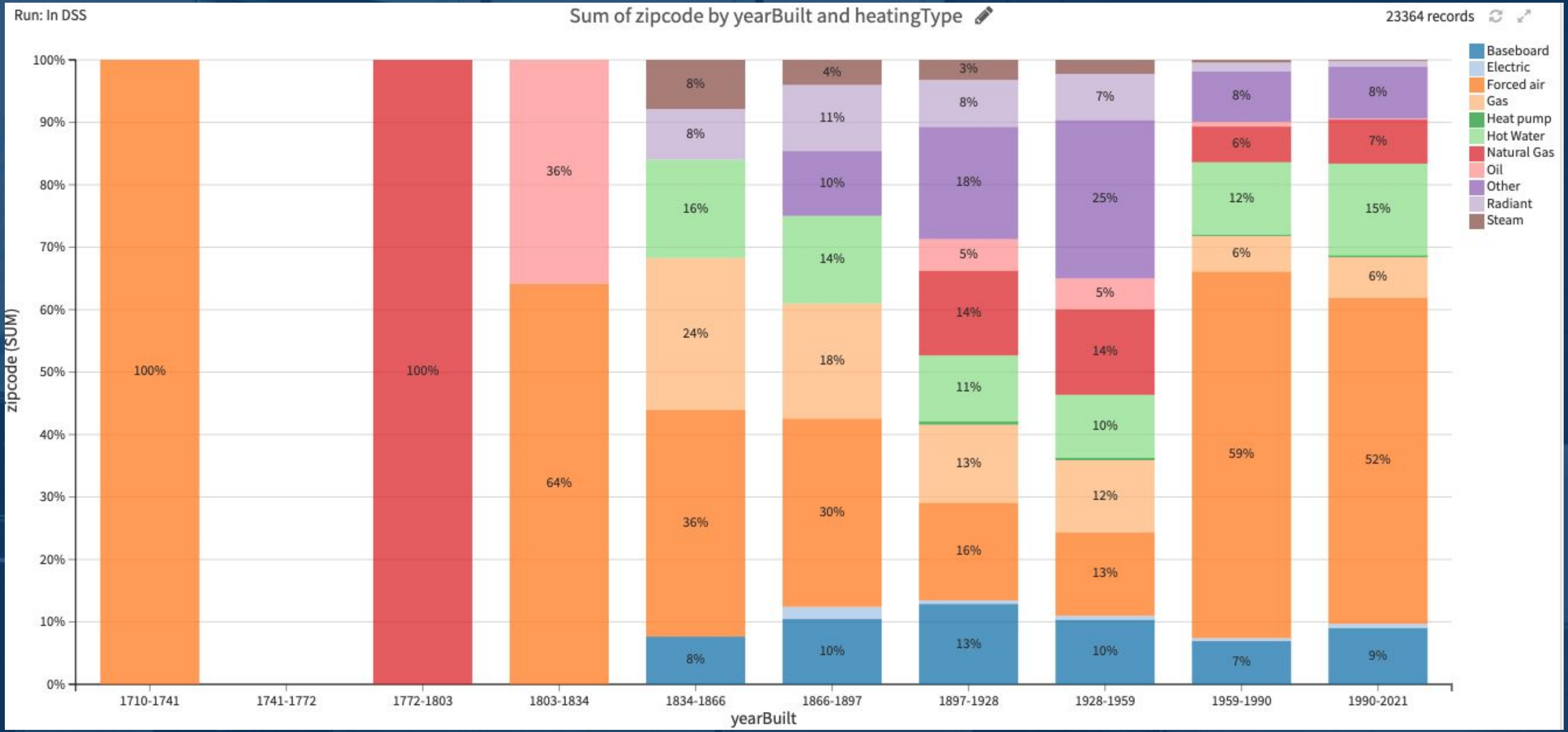
1. **The Dutch first settled along the Hudson River in 1624; two years later they established the colony of New Amsterdam on Manhattan Island. In 1664, the English took control of the area and renamed it New York.**

From which year, from which year rapid construction of building was started?



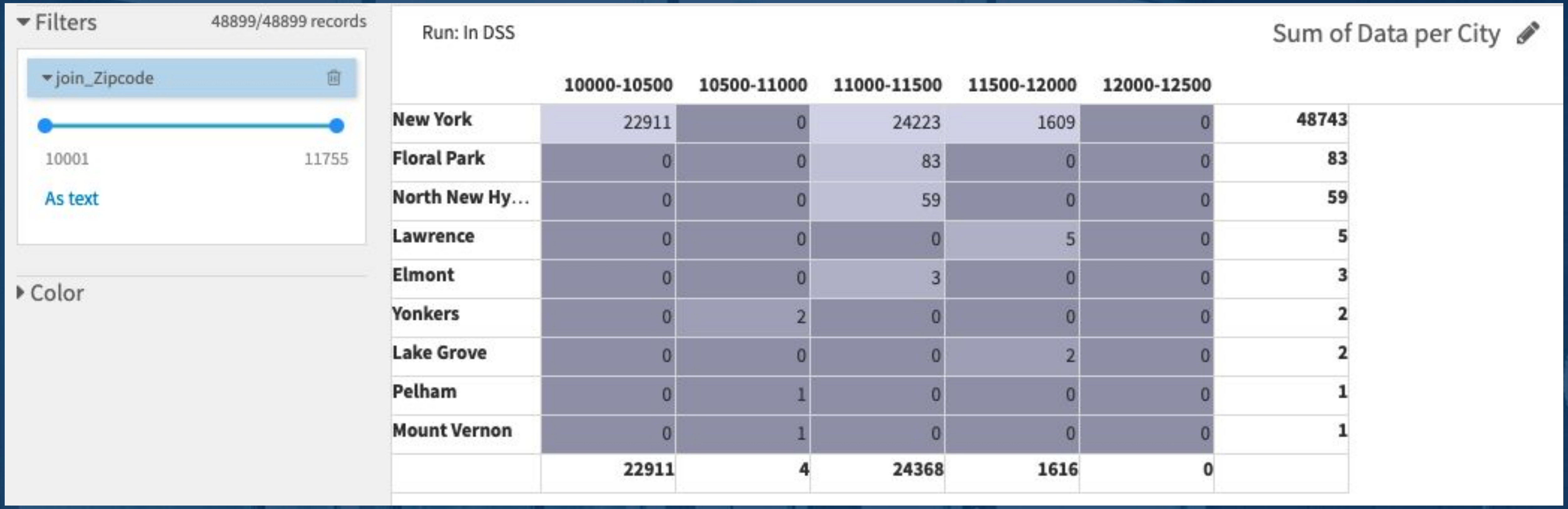
The construction of building started in late 1800s.

2. As per nyc.gov, the most commonly used heating system is Boiler (also known as Forced Air. Does the dataset reflect the same? If Yes, plot to prove it.



Yes, as you can see in the above graph, the most commonly used heating system is Forced Air

3. Clean the data with valid zip code, city & county and determine number of records per city & county.



Filters 48899/48899 records

join_Zipcode: 10001 – 11755

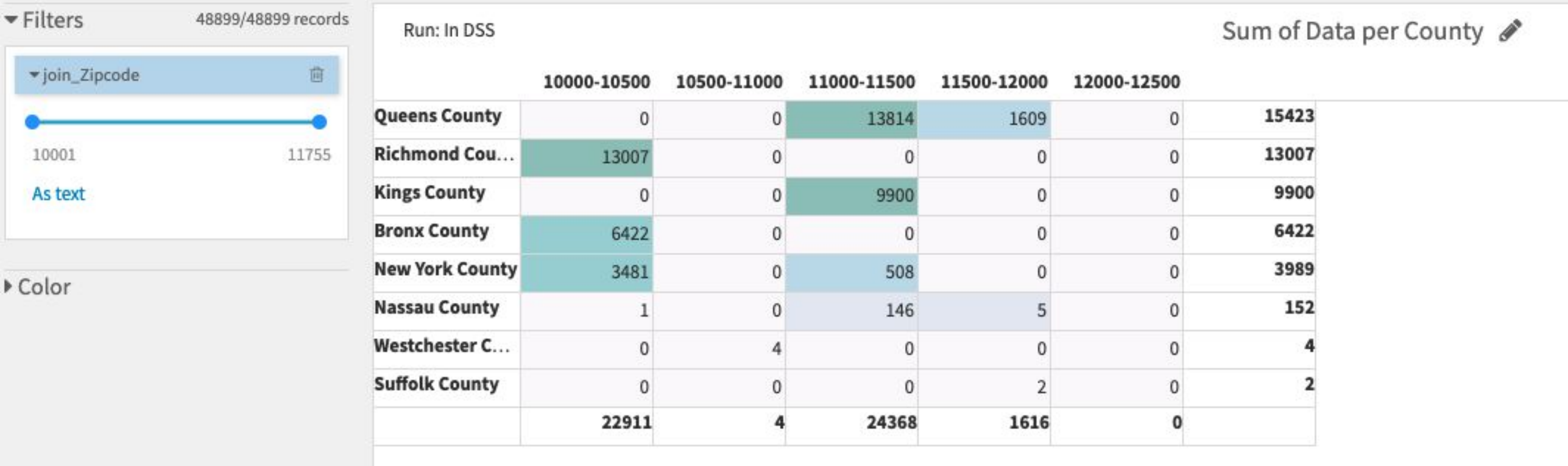
As text

Color

Run: In DSS

Sum of Data per City

	10000-10500	10500-11000	11000-11500	11500-12000	12000-12500	
New York	22911	0	24223	1609	0	48743
Floral Park	0	0	83	0	0	83
North New Hy...	0	0	59	0	0	59
Lawrence	0	0	0	5	0	5
Elmont	0	0	3	0	0	3
Yonkers	0	2	0	0	0	2
Lake Grove	0	0	0	2	0	2
Pelham	0	1	0	0	0	1
Mount Vernon	0	1	0	0	0	1
	22911	4	24368	1616	0	



Filters 48899/48899 records

join_Zipcode: 10001 – 11755

As text

Color

Run: In DSS

Sum of Data per County

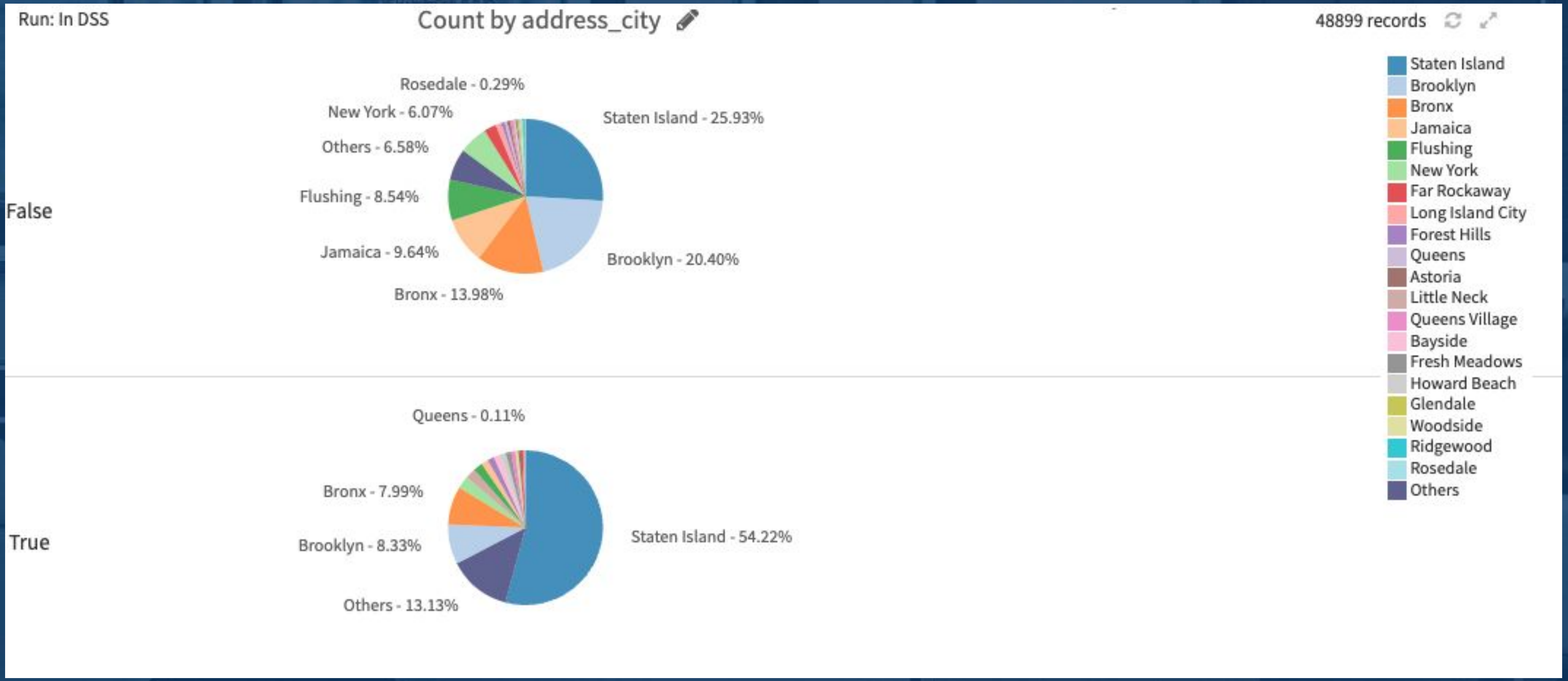
	10000-10500	10500-11000	11000-11500	11500-12000	12000-12500	
Queens County	0	0	13814	1609	0	15423
Richmond Cou...	13007	0	0	0	0	13007
Kings County	0	0	9900	0	0	9900
Bronx County	6422	0	0	0	0	6422
New York County	3481	0	508	0	0	3989
Nassau County	1	0	146	5	0	152
Westchester C...	0	4	0	0	0	4
Suffolk County	0	0	0	2	0	2
	22911	4	24368	1616	0	

With NYC opendata, county and city is arrived. Valid zip code range starts from 10001 to 11775, from which the above county & city was derived.

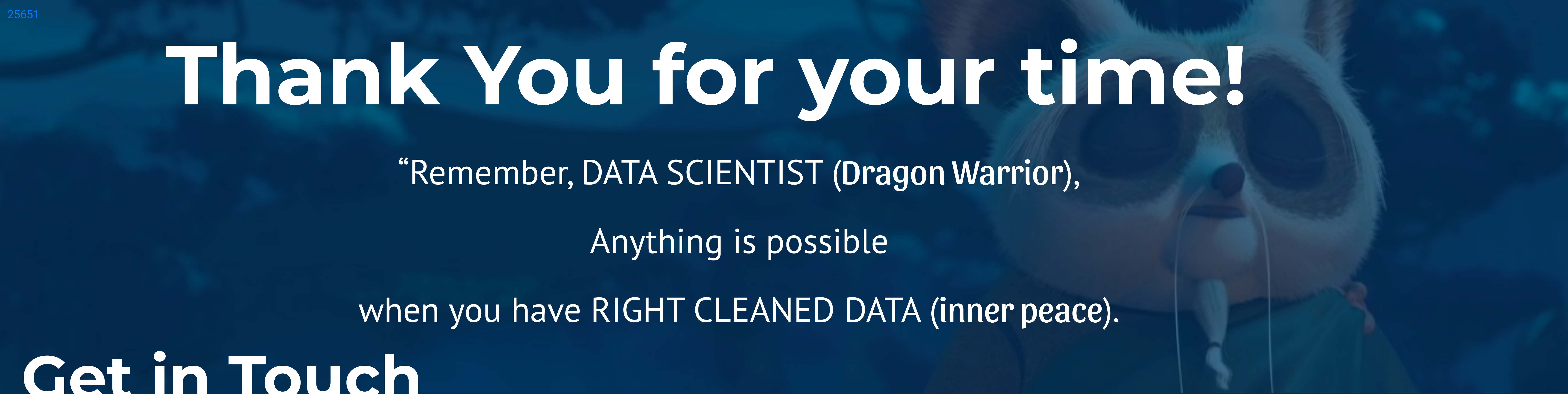
4. Buyer : I am moving from LA to New York, I am not able to find the right place.
Real Estate Agent : If you do not prefer the places I shared, do you have any special needs?

Buyer : Yes! I am pet lover and have 3 dogs and I cannot abandon them.
Real Estate Agent : Ok! Give me a minute!

Real Estate Agent : Contacts Revenue Revealers! (Hello Revenue Revealers)
Revenue Revealers : No worries! As per the data, he can move to Staten Island!



Real Estate Agent : Oh wow! Thanks for the info! This is so useful to convince my client!
Revenue Revealers : Feel free to ask questions any time and do not forgot to pay our share after sales!



Thank You for your time!

"Remember, DATA SCIENTIST (**Dragon Warrior**),

Anything is possible

when you have RIGHT CLEANED DATA (**inner peace**).

Get in Touch

Noubra Ashika

Ankit Shah

Mohammed Zeeshan Ali

Raviraj Ahire